

# Berkeley Water Center Eco-Science Infrastructure

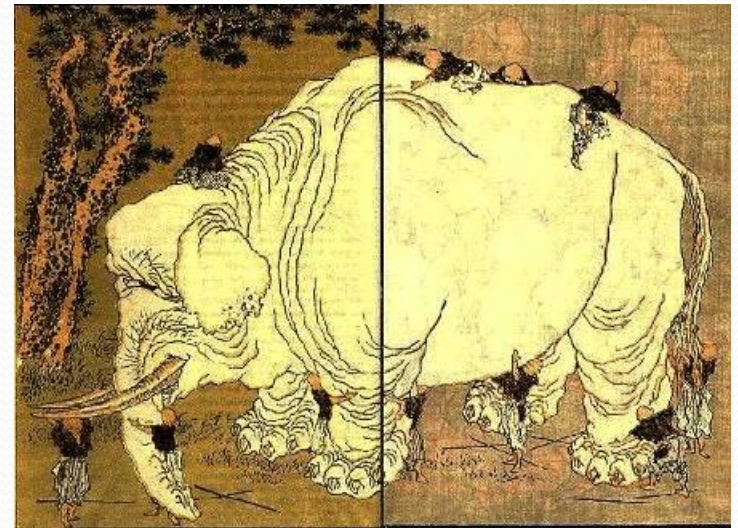
Deb Agarwal, Catharine van Ingen (Microsoft), Bora Beran (Microsoft), Monte Goode, Keith Jackson, and Robin Weber

# Ecological Data Avalanche/Landslide/Tsunami



The era of remote sensing, cheap ground-based sensors and web service access to agency repositories is here

- ▶ Extracting and deriving the data needed for the science remains problematic
  - ▶ Specialized knowledge
  - ▶ Finding the right needle in the haystack



# Data Availability is Often no Longer the Problem

- Broad range of agencies make data available via the web
  - Wide diversity of interfaces
  - Different data formats
  - Different representations
  - Interfaces difficult to navigate
  - Different periods of record
- Gaps in data and diverse periods of record
- Research groups typically use flat files on an internal server
- Data continually updated



# USGS NWIS Data Interface

The image displays four overlapping screenshots of the USGS NWIS Data Interface web application, illustrating its various components and search capabilities.

**Top Left Screenshot: USGS Surface-Water Data for the Nation - Mozilla Firefox**

This screenshot shows the main navigation page with the following sections:

- Real-time Data** (8,728 sites): Real-time data are time-series (15-minute intervals) data from automatic gauging stations that represent the most current measurements. Measurements are common minute intervals and transmit data every 1-4 hours. Data are available online for 31 days.
- Daily Data** (25,276 sites): Daily values are summarized for each day for the period of record. The daily mean, minimum, and/or other derived statistics are published, and more recent accuracy has not been verified.
- Statistics** (24,103 sites): Statistics are computed for time-series data at each site. Summaries of approved historical data are available for daily, monthly, and annual periods.

**Top Right Screenshot: USGS Surface-Water Daily Data for the Nation - Mozilla Firefox**

This screenshot shows the search interface with the following elements:

- USGS National Water Information System: Web Interface**
- USGS Water Resources**
- Choose Site Selection Criteria**
- Surface-water daily data example**
- Select sites which meet all**
- Define one or more values for each**
- Hydrologic Region** -- select one or more
  - 01 New England Region
  - 02 Mid-Atlantic Region
  - 03 South Atlantic-Gulf Region
  - 04 Great Lakes Region
  - 05 Ohio Region
  - 06 Tennessee Region
- Site type** -- select one or more
  - Meteorological
  - Ground Water
  - Spring
  - Estuary
  - Lake/Reservoir
  - Stream/River
- Available parameters** -- select sites
  - Select one or more parameters --
  - Water Level/Flow Parameters**
    - Depth to water level, ft below land
    - Direction of stream flow, magnetic
    - Elevation above NGVD 1929, ft
    - Elevation above NGVD 1929, m
    - Elevation of reservoir water surface
    - Flow rate, instantaneous, Mgal/d
    - Flow rate, instantaneous, gal/min
    - Gage height, above datum, m
    - Gage height, above datum, meter

**Bottom Left Screenshot: USGS Surface-Water Daily Data for the Nation - Mozilla Firefox**

This screenshot shows the search results page with the following elements:

- USGS National Water Information System: Web Interface**
- USGS Water Resources**
- Choose Site Selection Criteria**
- Surface-water daily data example**
- Select sites which meet all**
- Define one or more values for each**
- Hydrologic Region** -- select one or more
  - 01 New England Region
  - 02 Mid-Atlantic Region
  - 03 South Atlantic-Gulf Region
  - 04 Great Lakes Region
  - 05 Ohio Region
  - 06 Tennessee Region
- Site type** -- select one or more
  - Meteorological
  - Ground Water
  - Spring
  - Estuary
  - Lake/Reservoir
  - Stream/River
- Available parameters** -- select sites
  - Select one or more parameters --
  - Water Level/Flow Parameters**
    - Depth to water level, ft below land
    - Direction of stream flow, magnetic
    - Elevation above NGVD 1929, ft
    - Elevation above NGVD 1929, m
    - Elevation of reservoir water surface
    - Flow rate, instantaneous, Mgal/d
    - Flow rate, instantaneous, gal/min
    - Gage height, above datum, m
    - Gage height, above datum, meter

**Bottom Right Screenshot: USGS Surface-Water Daily Data for the Nation - Mozilla Firefox**

This screenshot shows the search results page with the following elements:

- USGS National Water Information System: Web Interface**
- USGS Water Resources**
- Choose Site Selection Criteria**
- Surface-water daily data example**
- Select sites which meet all**
- Define one or more values for each**
- Hydrologic Region** -- select one or more
  - 01 New England Region
  - 02 Mid-Atlantic Region
  - 03 South Atlantic-Gulf Region
  - 04 Great Lakes Region
  - 05 Ohio Region
  - 06 Tennessee Region
- Site type** -- select one or more
  - Meteorological
  - Ground Water
  - Spring
  - Estuary
  - Lake/Reservoir
  - Stream/River
- Available parameters** -- select sites
  - Select one or more parameters --
  - Water Level/Flow Parameters**
    - Depth to water level, ft below land
    - Direction of stream flow, magnetic
    - Elevation above NGVD 1929, ft
    - Elevation above NGVD 1929, m
    - Elevation of reservoir water surface
    - Flow rate, instantaneous, Mgal/d
    - Flow rate, instantaneous, gal/min
    - Gage height, above datum, m
    - Gage height, above datum, meter

**Search Results Summary:**

**Site Selection Results -- 2 sites found**

Parameter codes = 00618, 99060  
Site type = Surface Water  
State/Territory = California

Save file of selected sites to local disk for future upload

Data for individual sites can be obtained by selecting the site number below

Agency	Site Number	Site Name
USGS	10336610	UPPER TRUCKEE RIVER AT SOUTH LAKE TAHOE CALIF
USGS	10336780	TROUT CREEK NR. TAHOE VALLEY CALIF

**Questions about sites/data?**  
**Feedback on this web site**

[Top](#)  
[Explanation of terms](#)  
[Subscribe to NWISWeb notifications](#)

# EPA STORET

EPA > STORET > Stations by Geographic Location - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://iaspub.epa.gov/storpub/DW\_stationcriteria

Advanced Computing for Science dsd-admin Moderator Authentic... proposalCENTRAL

### Geographic Location

Select a single type of location search that you wish to perform (state/county, latitude/longitude, or drainage basin/HUC).

☒ State/County

State Name: CALIFORNIA County Name: SONOMA

☐ Latitude/Longitude (in decimal degrees)

West Limit: 180

☐ Drainage Basin/HUC

Cataloging Unit: ALL [Look Up]

### Station Type

Select one or more Station Type(s).

PRIMARY_TYPE	SECONDARY_TYPE
Select All	
River/Stream	
Lake	
Great Lake	
Well	
Facility Industrial	
Facility Municipal Sewage (POTW)	
Facility Other	
Facility Privately Owned Non-indus	
Facility Public Water Supply (PWS)	

### Characteristic

Find Stations with at least one reported value of the selected characteristic.

Use the Characteristic Search to create a list of up to 100 stations.

Done

EPA > STORET > Station Search Summary - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://iaspub.epa.gov/storpub/DW\_STATION\_COUNT

Advanced Computing for Science dsd-admin Moderator Authentic... proposalCENTRAL EPA > STORET > Stations by G... EPA > STORET > Station S...

### Number of Stations Returned: 41

#### Search Parameter Values

State: CALIFORNIA

County: SONOMA

Station Type(s): River/Stream

Characteristic(s): ALL

Organization(s): ALL

Select 'Back' to modify search parameters and refine your query.  
Select Batch Processing to generate the report offline based on your current selections.  
You may customize the content of your report by selecting Data Elements below.

Please enter your email address:

Please specify three characters to prefix your report name:

#### Batch Processing

Note: Download files will exceed maximum limits for spread sheet applications (ie. EXCEL - Worksheet size: 65,536 Rows.)

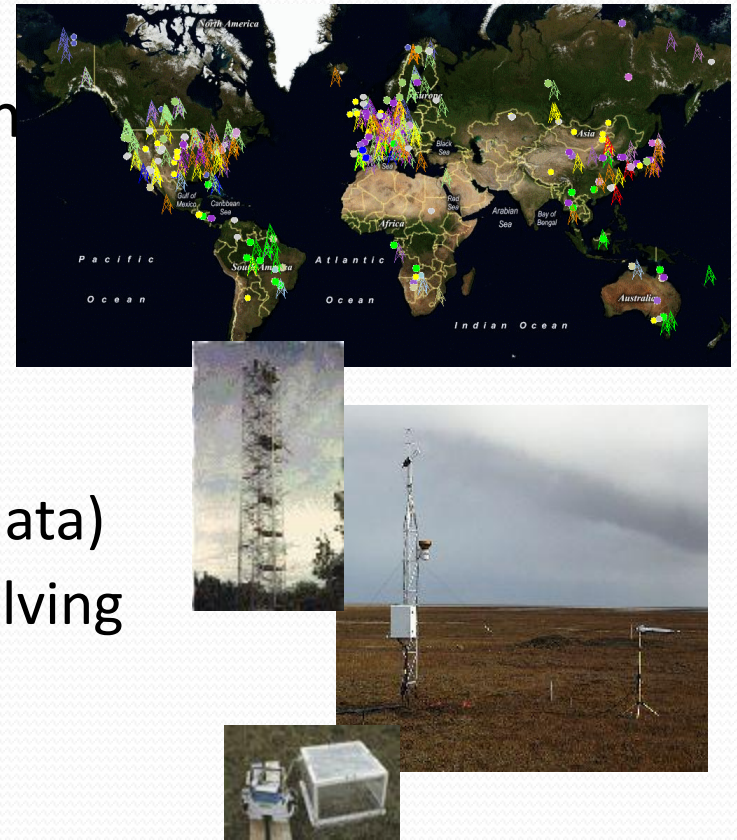
#### Select Data Elements for Report

<input checked="" type="checkbox"/> Org ID	<input type="checkbox"/> Elevation (w/ Units)
<input checked="" type="checkbox"/> Beach ID	<input type="checkbox"/> Additional Elevation Info*
<input checked="" type="checkbox"/> Station ID	<input type="checkbox"/> Country Name
<input checked="" type="checkbox"/> Station Name	<input checked="" type="checkbox"/> State
<input type="checkbox"/> Org Name	<input checked="" type="checkbox"/> County
<input checked="" type="checkbox"/> Primary Type	<input type="checkbox"/> Hydrologic Unit Code
<input type="checkbox"/> Secondary Type	<input type="checkbox"/> Hydrologic Unit Name

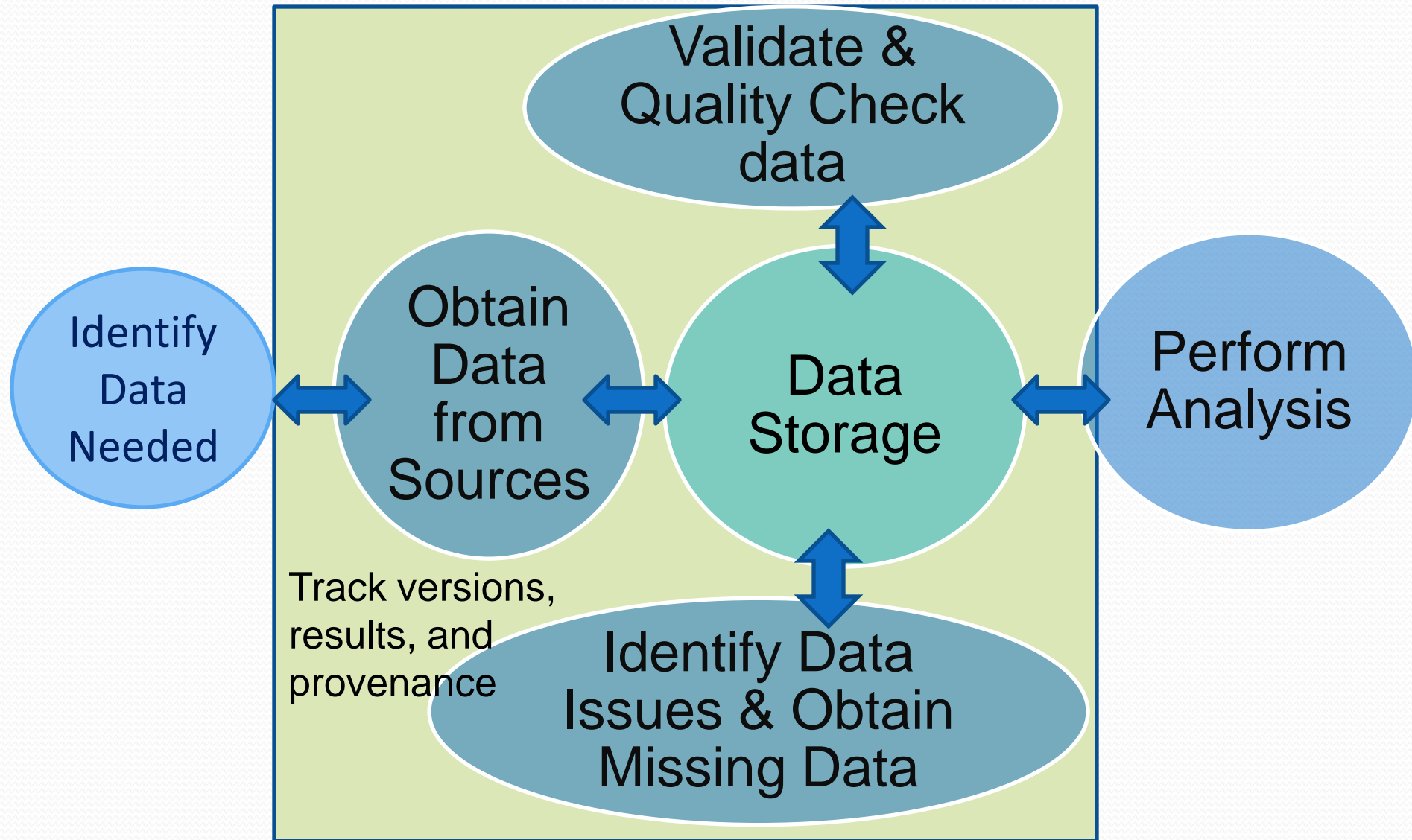
Done

# Example Experiment Dataset - FLUXNET

- Global carbon flux and micrometeorological measurements
  - 960 site years of data
    - 40 variables and 60 quality tags
    - 1/2 hourly measurements
  - 250 sites
  - 160 site investigators (collecting data)
  - 65 synthesis paper proposals involving ~130 collaborators
- Goal: Understand the role of photosynthesis in carbon cycle



# Target Data Analysis Flow





# Developing Common Eco-Science Data Infrastructure

- Graphical data selection tool
- Data ingested into a database
- Database schema organized by dataset, site, time, and variable
- On-line analytical processing (OLAP) data cubes provide browsing access
- Reports provide web pages describing the data
- Collaboration tools facilitate communication between data providers and data users as well as between data users

[illegible]




## LAYERS

- Watershed    
- Ecoregion    
- Geology    

## SEARCH

### DATA

- Geographic area   
- Time Frame  
- Keyword

discharge



SUBMIT

RESET



**Site Name :** PETALUMA R A COPLAND PUMPING STATION A PETALUMA

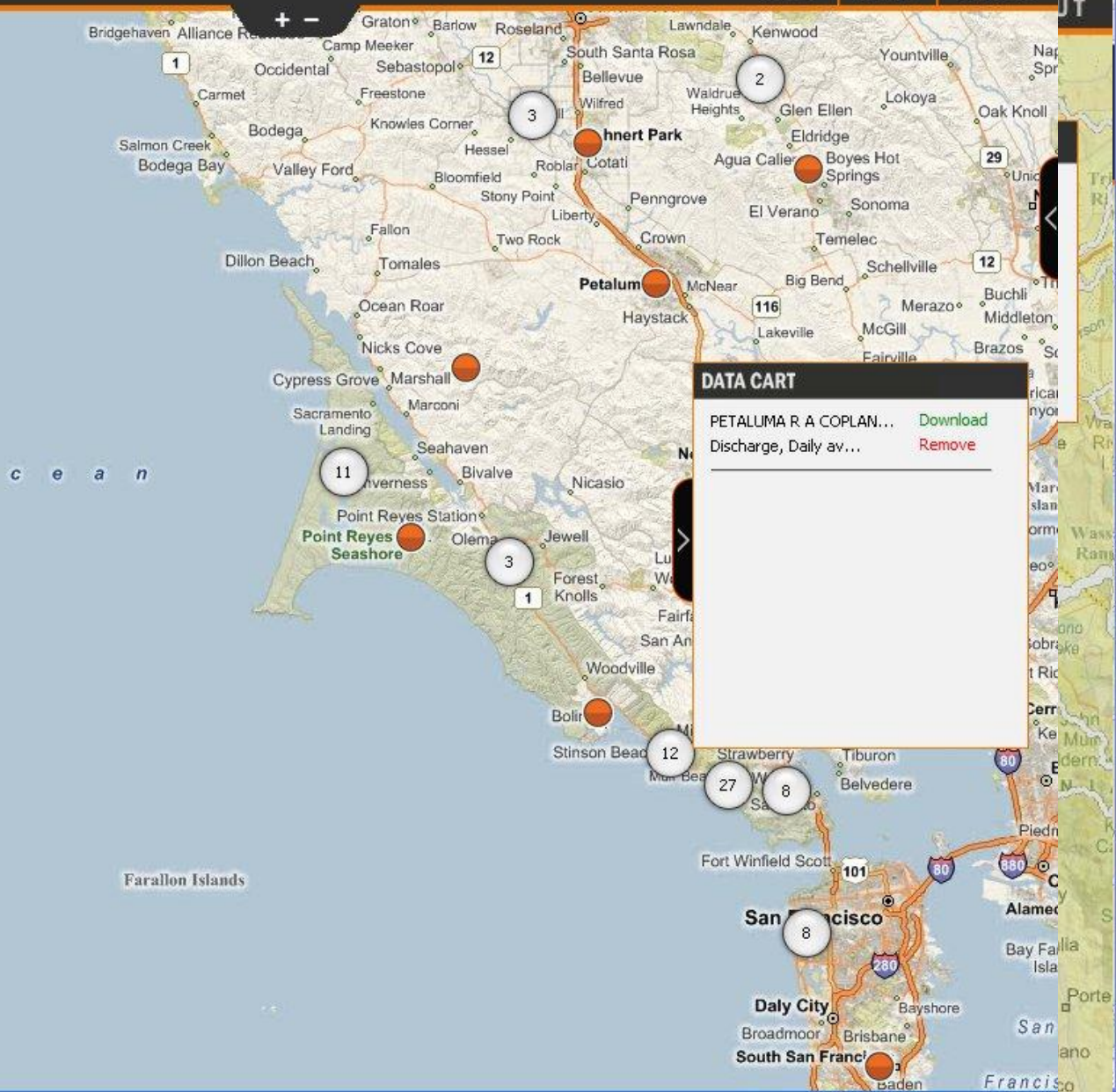
**Operated by :** National Water Information System

### Available Measurements :

Discharge, Daily average  

### Relevant Measurements :

Discharge, Daily average  



## DATA CART

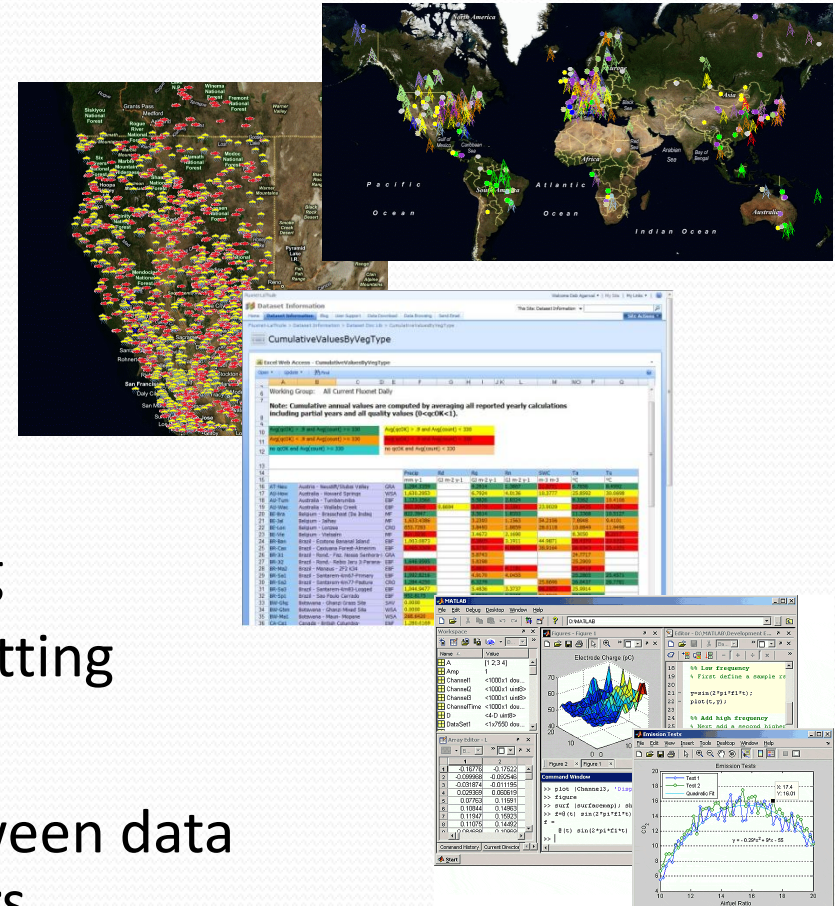
PETALUMA R A COPLAND PUMPING STATION A PETALUMA  
Discharge, Daily average [Download](#) [Remove](#)

# Behind SciScope

- Contains information on where, when (time frame) and what is being measured for about 1.65 million sites in the US
- Scraped/crawled on a regular basis
- SciScope currently hosts only metadata.
- Data are requested on the fly from the original publisher using web service wrappers written specifically for each data source.
- Data are reformatted to provide a unified view over the repositories.

# Interaction and Visualization of Data

- Mashup maps
  - locate sensors
- Data summary reports
  - Data browsing
  - Data selection
- OLAP data cube interfaces
  - Matlab – analysis and plotting
  - Excel – data browsing and plotting
- Collaboration capabilities
  - Improve communication between data producers and data consumers





# Pivot Tables

Excel PivotTable interface showing a PivotTable with the following settings:

- Exdatumtype:** (none)
- Repeat:** 0
- Datumtype:** Rg
- Value units:** (Multiple Items)
- Dataset:** June download from ORNL web site

The PivotTable is set to show **MaximumValue** for the **Site** and **Latitude** fields. The data is organized by **Yy** (Year) and **Yymm** (Month).

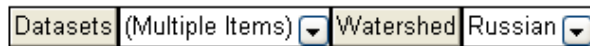
Yy	Yymm	Yymdd	Site	Latitude
2001	2001-09	2001-09-01	Brazil -- Tapajos (Santarem,Km83), Logged Forest	Canada -- BOREAS NSA - 1930 burn site
		2001-09-02		Canada -- BOF
		2001-09-03		
		2001-09-04		
		2001-09-05		414.6756592
		2001-09-06		557.3077393
		2001-09-07		
		2001-09-08		
		2001-09-09		
		2001-09-10		
		2001-09-11		
		2001-09-12		
		2001-09-13		
		2001-09-14		
		2001-09-15		
		2001-09-16		
		2001-09-17		
		2001-09-18		
		2001-09-19		
		2001-09-20		
		2001-09-21		
		2001-09-22		
		2001-09-23		

The PivotTable Field List on the right shows the following fields:

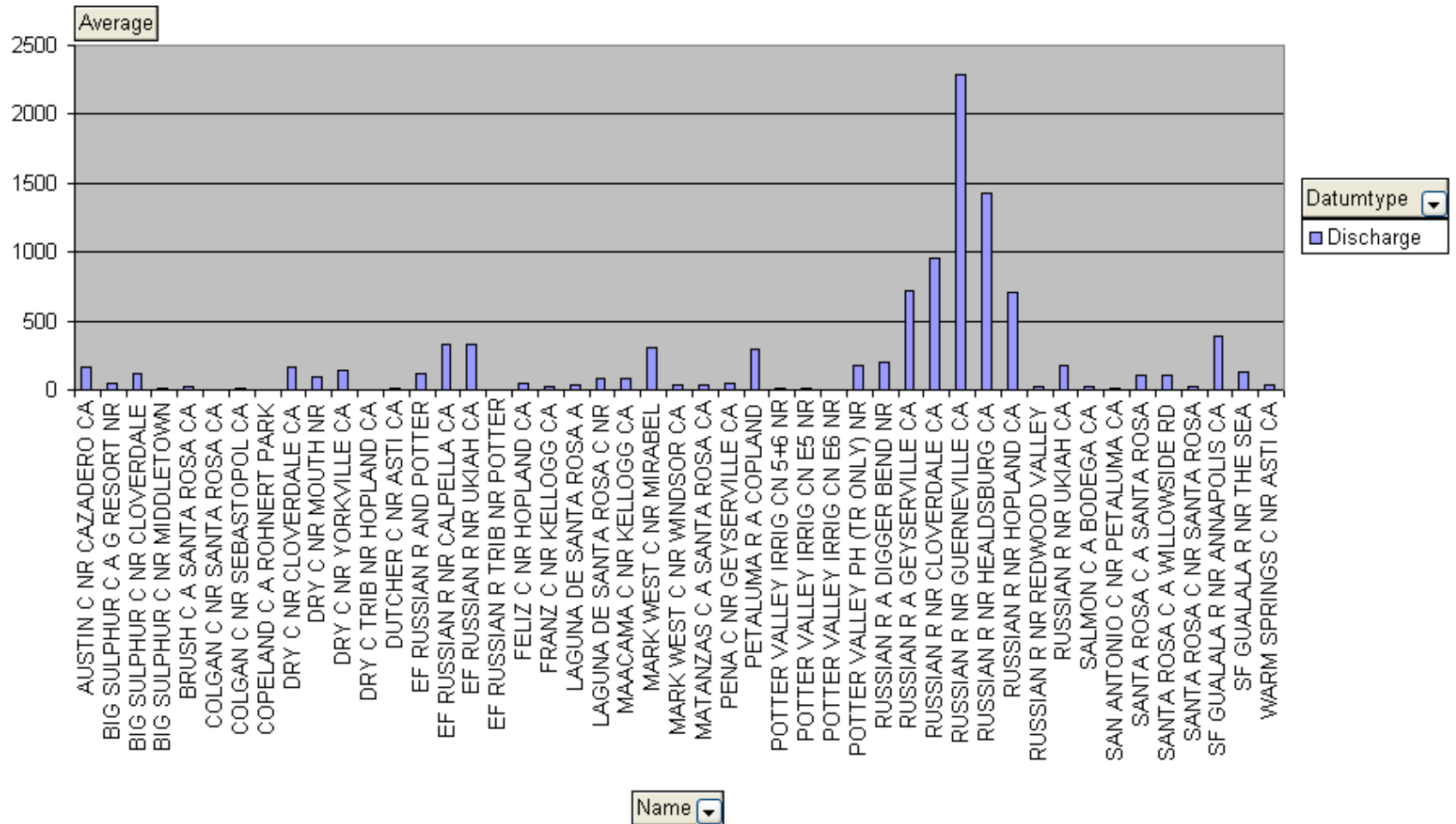
- Dataset
- Datumtype
- Name
- Exdatumtype
- Value units
- Repeat
- Ecosystem
- Latitude
- Site
- Day of Year
- Hour of Day



# Pivot Charts



Discharge



# MatLab Interface

**Berkeley Water Center Data Cube Query Builder** Release 2008-05-21

Cube: LatestAmfluxL3L4Daily

**Dimensions**

**Datatype** Non-Empty

Datatype ☒

Exdatatype ☒

Offset ☒

Quality ☒

Site IGBP To Site-IGBP... ☒

Timeline Year ☒

WorkingGroup ☒

**Filters**

Time	Datum Type	Site	Other
<span>Year</span>	<span>Datatype</span>	<span>Country</span>	
1991 1992 1993 1994 1995	GPP H H2O LE NEE	Brazil Canada USA	
<span>Month Of Year</span>			
4 5 6 7 8			

**Measure**

YearlyCalc

**Code**

```

cube_query_out =
QueryCube('https://bwc.berkeley.edu/mdxconnect2/Default.aspx?db=', 'LatestAmfluxL3
L4Daily', 'LatestAmfluxL3L4Daily', username, password, '[Measures].[YearlyCalc]', '[Data
tmtype].[Site].[Timeline]', '[Datumtype].[Datumtype]' [Site].[IGBP To
Site].[IGBPClass]' [Timeline].[Year].[Year]', [1 1 1

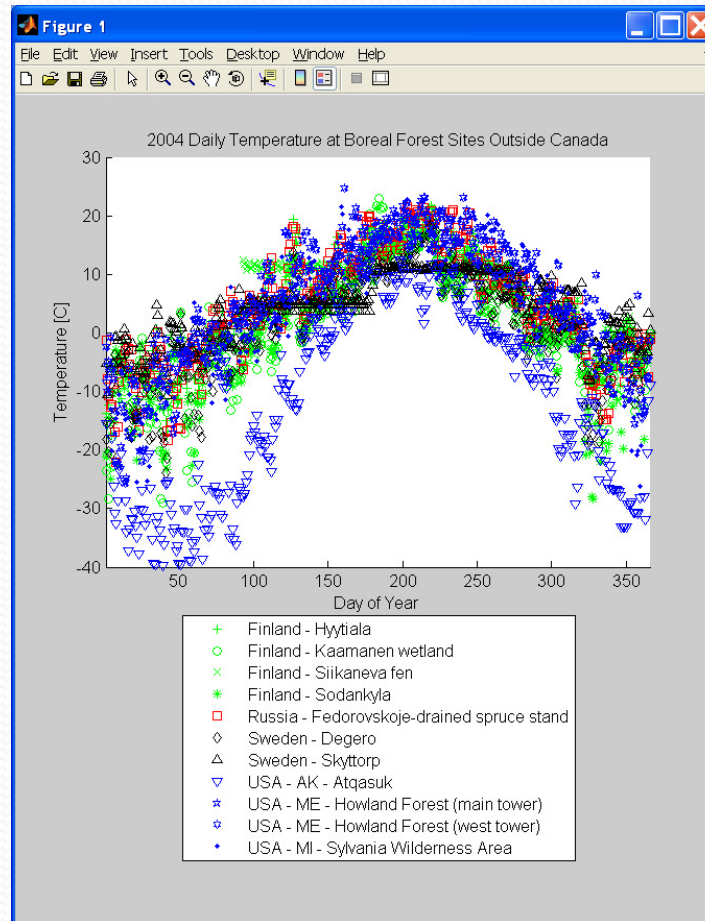
```

**Submit Query**

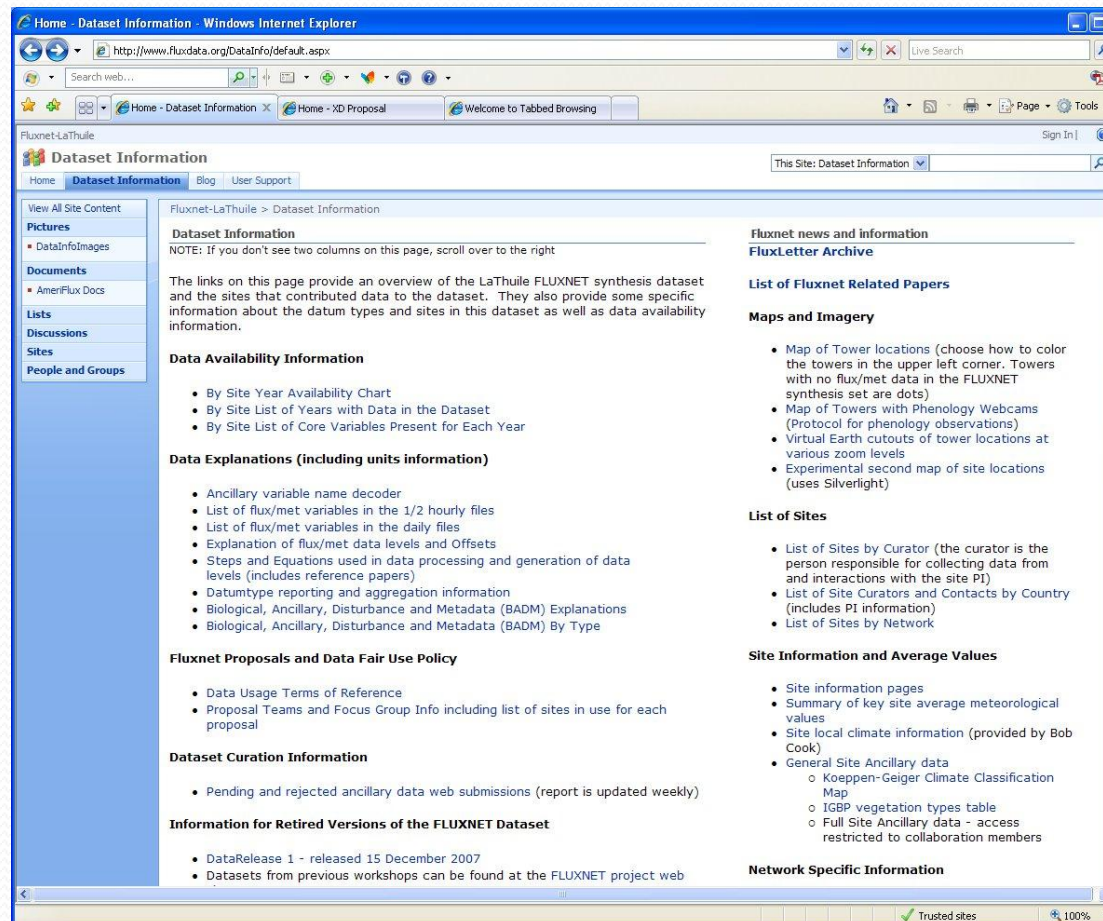
into variable

cube\_query\_out

# Plotting Results



# SharePoint Portal





# SharePoint Hosted Reports

Dataset Doc Lib - CumulativeValuesByVegType - Windows Internet Explorer

http://www.fluxdata.org/DataInfo/Dataset%20Doc%20Lib/CumulativeValuesByVegType.aspx

Search web...

Dataset Doc Lib - Cumula... X Home - XD Proposal Welcome to Tabbed Browsing

Fluxnet-LaThuile Sign In

Dataset Information

Home Dataset Information Blog User Support

Fluxnet-LaThuile > Dataset Information > Dataset Doc Lib > CumulativeValuesByVegType

CumulativeValuesByVegType

Excel Web Access - CumulativeValuesByVegType

Open Update Find

A B C D E F G H I J K L M N

1 Cumulative Annual Variable Values by Site

2 Download Excel

3 Generated: 9/2/2008 5:30:13 PM

4

5 Working Group: All Current Fluxnet Daily

6

7 Data Release 2+BADM (preliminary)

8 Data as of 2008-02-16

9 Note: Cumulative annual values are computed by averaging all reported yearly calculations including partial years and all quality values (0 < qcOK < 1).

10

11 Avg(qcOK) > .9 and Avg(count) >= 330 Avg(qcOK) > .9 and Avg(count) < 330

12 Avg(qcOK) < .9 and Avg(count) >= 330 Avg(qcOK) < .9 and Avg(count) < 330

13 no qcOK and Avg(count) >= 330 no qcOK and Avg(count) < 330

14

15

16

17

			Precip mm y-1	Rd GJ m-2 y-1	Rq GJ m-2 y-1	Rn GJ m-2 y-1	SWC m-3 m-3	Ta °C	Ts °C
18	AT-Neu	Austria - Neustift/Stubai Valley	GRA	1,284.3359	4.2914	1.3867	31.8751	6.7856	8.4992
19	AU-Fog	Australia - Fogg Dam	SAV	1,463.1003	7.3308	4.5664	96.0497	25.6354	25.3748
20	AU-How	Australia - Howard Springs	WSA	1,630.2953	6.7924	4.0136	10.3772	25.8592	30.0698
21	AU-Tum	Australia - Tumbarumba	EBF	1,123.3560	5.3826	2.8324		9.3362	10.4106
22	AU-Wac	Australia - Wallaby Creek	EBF	870.0000	0.6696	1.4264	25.8877	14.4618	9.0123
23	BE-Bra	Belgium - Bresschaat (De Inslag Forest)	MF	822.3947		3.5614	1.8303	11.3368	10.5127
24	BE-Jal	Belgium - Jalhay	MF	1,633.4386	3.2393	1.1563	54.2106	7.8948	9.4101
25	BE-Lon	Belgium - Lonze	CRO	653.7283	3.8493	1.8859	28.0118	10.8849	11.9498
26	BE-Vie	Belgium - Vielsalm	MF	821.0230	3.4672	2.1690		8.3050	8.2217
27	BR-Ban	Brazil - Ecotone Bananal Island	EBF	1,003.0873	2.2000	3.1911	44.9871	26.1478	23.0000
28	BR-Cax	Brazil - Caxiua Forest-Almeirim	EBF	1,465.3309	5.5755	4.8808	10.5144	26.0343	25.1231
29	BR-Jil	Brazil - Rond.-Faz. Nossa Senhora-i	GRA		5.8743			24.7717	

Trusted sites 100%

# Acknowledgements

Berkeley Water Center, University of  
California, Berkeley, Lawrence  
Berkeley Laboratory

Jim Hunt  
Deb Agarwal  
Robin Weber  
Monte Good  
Rebecca Leonardson (student)  
Matt Rodriguez  
Carolyn Remick  
Susan Hubbard

University of Virginia  
Marty Humphrey  
Norm Beekwilder

Microsoft

Catharine van Ingen  
Bora Beran  
Jayant Gupchup (student)  
Savas Parastidis  
Andy Sterland  
Nolan Li (student)  
Tony Hey  
Dan Fay  
Jing De Jong-Chen  
Stuart Ozer  
SQL product team  
Jim Gray

Ameriflux Collaboration

Beverly Law  
Youngryel Ryu (postdoc)  
Tara Stiefl (student)  
Gretchen Miller (student)  
Mattias Falk  
Tom Boden  
Bruce Wilson

Fluxnet Collaboration

Dennis Baldocchi  
Rodrigo Vargas (postdoc)  
Dario Papale  
Markus Reichstein  
Bob Cook  
Susan Holladay  
Dorothea Frank

National Marine Fisheries Service  
Sonoma Ecology Center

The saga continues at  
<http://bwc.berkeley.edu> and  
<http://www.fluxdata.org>